4-26-2020

Comparison of Statistical Models for Imputation of Missing Data in Clinical Trials

Vaishnavi Purandare
vjp1640@rit.edu

Rochester Institute of Technology

Master's Thesis

Comparison of Statistical Models for Imputation of Missing Data in Clinical Trials

Author:
Vaishnavi Purandare

Supervisor:
Dr. Robert Parody

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

in

Applied Statistics
College of Science
Department of Mathematical Sciences

April 26, 2020

Rochester Institute of Technology

Abstract

Dr. Robert Parody
School of Mathematical Sciences

Master of Science

Comparison of Statistical Models for Imputation of Missing Data in Clinical Trials

By Vaishnavi Purandare

Missing data is an integral part of clinical trials and its analysis. This study discusses the downsides of having missing values in clinical data, traditional methods used to resolve this issue and some techniques which can be implemented to remedy the same.

The data used for the study is simulated from Theophylline data from Pinheiro and Bates (1995). The simulated data measures the Theophylline drug concentration in the body of 10 Subjects over 24 hours. There are three cases considered with increasing number of randomly created missing values for the Concentration variable. Subsets are created to fit linear and quadratic linear subject and population models. The fitted models are compared using Sums of Squares of Imputation and Imputed R^2. These comparative techniques indicate that replacing the missing values in clinical data with appropriate estimates, while maintaining the authenticity of the data, is feasible.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my Graduate research advisor, Dr. Robert Parody for inspiring this study and guiding me along the way towards its completion and success. His continued support and encouragement throughout the Master's program are highly appreciated.

Besides my advisor, I wish to thank the rest of my thesis committee members, Dr. Carol Marchetti and Dr. Linlin Chen for their efforts. Dr. Marchetti's advice has been monumental in finalizing the study and has brought out the best in me.

I would like to thank all the professors and staff of the Applied Statistics program for making my time at RIT an insightful and memorable experience. My special thanks to Shawna Hayes for her diligent efforts in making sure I stay on top of things.

Lastly, I would like to thank my parents, family, and friends for their unwavering belief in me. Their encouraging words have helped me through the tough times. My heartfelt thanks.

Thesis Committee Approval

Dr. Robert Parody Date

Thesis Advisor / Associate Professor / School of Mathematical Sciences

---

Dr. Carol Marchetti Date

Committee Member / Professor / School of Mathematical Sciences

---

Dr. Linlin Chen Date

Committee Member / Associate Professor / School of Mathematical Sciences

Table of Contents

List of Tables

List of Figures

Section 1

Introduction

1.1 Background and Motivation

Missing data is an integral part of clinical trials. It is extremely rare to get clinical trial data without any missing values. A variable from data can be considered as missing if the value of the variable, may it be the response variable or one of the covariates, for a participant is not recorded.

The type of the study determines what kind of missing data you can expect. A longitudinal study might have points missing from a subject's repeated measures either at the beginning of the trial due to difficulty in measuring the outcome at an early stage or at the end of the trial because of dropouts. Omitting the subjects with the missing data from the trial can greatly impact the analysis. Losing subjects from a study will result in loss of information. It will also lead to reduced power due to a smaller sample size.

Few instances of what can be considered missing from a certain study are missing mortality rate from a survival study; missing blood group, TSH levels etc. in laboratory studies, diameter, size of a tumor in an oncology study, missing efficacy for a pre-clinical trial, missing half-life value for a drug study, etc.

It is significant that we check the effect of the missing covariate and response data on the analysis of the study. The impact of the missing data might vary depending on the objectives of the study.

There are various methods used to treat these missing points. The methods widely in practice include Complete Case Analysis, minimum-maximum substitution, average substitution, missing indicator method, (Cohen and Cohen (1983) and Cohen, Cohen, West and Aiken (2003)), etc. The most common method to analyze a data with missing points, outcome or co-variate, is to omit the subjects with any missing data. This method is known as the Complete Case Analysis. It is a widely used procedure as it is a default for quite a lot of analyses in statistical analysis packages such as SAS, STATA and R. For instance, SAS will by default ignore the observations with missing values while fitting a general linearized model using 'proc glm' for a given dataset.

In a lot of cases, the Complete Case Analysis is not the appropriate method. It may lead to misleading conclusions. We would also have less power for assessment of treatment effects and significance of covariates. The degrees of freedom available for prediction of the treatment means also get reduced.

Other methods include substituting the missing values with average, minimum or maximum values of the variable. These methods have their own pros and cons which will be discussed in detail later.

To remedy this issue of missing data in clinical trials, we can build models to predict the missing values. The predicted values of the missing points can be incorporated in the dataset for the analysis. This is termed as imputation of missing data. This will give us better insights as a result of increased power and sample size. We need statistical measures to compare these models to find the one that is best for data imputation. This study incorporates the Sums of Squares of Imputation and the Imputed R^2, calculated using the Sums of Squares of Imputation, as the measures for assessing the models used for imputation.

1.2. Study Objectives

The objectives of my thesis are as follows:

1. Choose the best model among the models built for data imputation for simulated clinical trial missing data
2. Compare the Sums of Squares of Imputation for all of the models
3. Validate that the SSI and the Imputed R^2 are measures which can be used for comparison of the models

1.3 Study Layout

The next section gives a brief introduction of Clinical Trials and the common terms used in Clinical Research. It gives a theoretical background of the models used in the study. It also introduces the statistical measure used to compare the models to find the best model for prediction.

Section 3 focuses on the procurement of the simulated data used for this study. It also gives a brief summary of the steps taken to incorporate missing points in the data followed by summary of the statistical models built for predicting the missing data points. Section 4 includes the literature review done for this study. It briefly explains the other strategies used to deal with missing data in clinical trials.

Section 5 includes the descriptive statistics of the study, the outputs from the models and the in-depth comparison of the models using goodness of fit measures to find the best model for imputation. The final section provides the final results of the comparative study along with the conclusion of the research. It discusses regarding the shortcomings of the methods used and scope for taking this research further.

Section 2

Background

2.1 Clinical Trials Summary

World Health Organization (WHO) defines a clinical trial as 'any research study that prospectively assigns human participants or groups of humans to one or more health related interventions to evaluate the effects on health outcomes. Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiological procedures, devices, behavioral treatments, process-of-care changes, preventive care, etc.'

In other words, any experiment, involving humans as subjects, conducted to check the safety or efficacy of drugs, medical procedures, medical devices, etc. can be considered as a Clinical Trial.

A clinical trial is conducted over four phases and is closely regulated by the Food and Drug Association (FDA).

Phase I - The scientist introduces an experimental treatment, drug or procedure, to a small group of subjects. The objectives of this phase are to check the safety of the treatment, decide the safe dosage, and record any side effects observed.

Phase II - The drug is now given to a larger group of subjects. The objectives of this phase are to determine the efficacy of the experimental treatment and monitor the side effects.

Phase III - The experimental treatment is now given to big groups of subjects. There might be more than one group of subjects. Here, the treatment group receives the treatment, while the control group receives the placebo (inactive or old treatment). The objectives of this phase are to validate the efficacy of the new treatment usually by comparison with old or commonly used treatments, monitor the side effects on the wide range of subjects. The crucial part of this Phase is that the results of the trial are submitted to FDA for the approval of the experimental treatment.

Phase IV - Post approval from the FDA, the new treatment is introduced in the market. The objective of this phase is to check long term effects of the treatment and monitor any unusual and unprecedented side-effects.

2.2 Keywords used in Clinical Trials

2.2.1. Outcome/ Response

The response variable or the variable of interest recorded for each subject of the clinical trial is called as the Outcome/ Response.

The responses may defer based on the type of trial. For instance, level of pain will be the outcome for an arthritis drug trial, iron concentration in body for a Haemoglobin supplement study, Chest X-ray for Tuberculosis Treatment Trial etc.

In the simulated data for this study, the observed 'Concentration of drug Theophylline in the body' will be the response variable.

2.2.2. Fixed Effect

Fixed effects are variables that are constant across subjects. It means that the effect of these variables is going to be the same regardless of the subject. We are usually interested in the effect of the fixed effect variable on the response variable.

In the simulated data for this study, the variable 'Time' will be the fixed effect.

2.2.3. Random Effect

Random effects are variables that vary across subjects. We are usually interested in the variability created in the response variable by the random effect variable rather than the effect of the random effect on the response.

In the simulated data for this study, the variable 'Subject' will be the fixed effect.

2.2.4. Linear Mixed Effect Model

Linear Mixed effects model is an extension of multiple linear regression where one of the independent variables is a random effect. There can be more than one fixed and random effect in the data. Typically, in clinical trial analyses, variables such as Drug concentration, Time, Weight, Age, Gender, etc. are fixed variables and the Subject or Patient is considered as the random effect.

The subjects are selected randomly from a larger pool of patients. We are interested in the whole population of subjects rather than the specific subjects selected for the trial. Each subject might react differently to a treatment and each treatment might affect the subject differently. We are interested in the variability of response caused by the subjects rather than the effect itself.

The Algebraic Form of Linear Mixed Effect Model is given below:

$$Yi = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \text{ (fixed)} + u_{1i} Z_{i1} + \ldots + u_{qi} Z_{iq} + i \text{ (random)}$$

where the index, i, is used to index subjects. Note that $Z_{i1}, \ldots, Z_{iq}$ is associated with the random effects $u_{1i}, \ldots, u_{qi}$ that are specific to subject i.

Since the LMM can take on one or more fixed or random effects, it is best to use the matrix form of this equation.

The model can be represented in Matrix form as below:

$Y_i = X_i\beta$ (fixed)$+Z_iu_i + i$ (random)

where $u_i \sim N(0, D)$ and $i \sim N(0, R_i)$

and D represents variance of u_i and R_i represents variance of i

2.2.5. Sums of Squares of Imputation

A summary measure called Sums of Squares of Imputation (SSI) is defined as the sum of squared differences of observed and expected response for the non-imputed data points.

SSI = $\sum_{i=1}^{n}(y'_i - \widehat{y'_i})^2$

Where, y'_i are the non-imputed data points. In other words, y'_i is the response of subjects without any missing data.

According to the definition given by Richardson-Harman and Parody (2016), the Sums of squares of imputation is the sum of the squared differences between the measurements (imputed and detectible) and the predicted values from the model fit.

A low SSI indicates a better data imputation and model fit.

This will be further discussed in Section 3 and Section 4

2.2.6. Imputed R^2

Imputed R^2 is calculated using Sums of Squares of Imputation as the numerator and the Total Sums of Squares as the denominator. Here, SSM denotes the Sums of Squares of Model.

Imputed $R^2 = 1 - \frac{SSI}{SST} = \frac{SSM}{SST}$

Imputed R^2 is the R^2 of the model after imputing the predicted values in the data.

The value of the measure falls between 0 and 1. The value of the Imputed R^2 ideally, should be close to 1.

A higher Imputed R^2 value indicated a good model for imputation.

Section 3

Methodology

3.1 Procurement of data using simulation

This section discusses how the data used for the comparative techniques has been simulated. The foundation of the simulated data is based on Theophylline data from Pinheiro and Bates (1995). The original data consists of serum concentrations of the drug Theophylline measured over a 25-hour period after oral administration for 12 subjects. Pinheiro and Bates (1995) have considered the first order one-compartment model for analyzing the data. One-compartment model is the simplest model that can be used to explain absorption and elimination of drug in a body. The model assumes that the body is a single uniform compartment and the drug is distributed instantaneously in the body. The first order one-compartment model is given below:

$$C_{it} = \frac{D k_{e_i} k_{a_i}}{Cl_i(k_{a_i} - k_{e_i})}\left[\exp(-k_{e_i}t) - \exp(-k_{a_i}t)\right] + e_{it}$$

where, C_{it} is the observed concentration of the i^{th} subject at time *t*, *D* is the dose of theophylline, k_{e_i} is the elimination rate constant for subject *i*, k_{a_i} is the absorption rate constant for subject *i*, Cl_i is the clearance for subject *i*, and e_{it} are normal errors. To allow for random variability between subjects, they assume

$$\begin{aligned} Cl_i &= exp(\beta_1 + b_{i1}) \\ ka_i &= exp(\beta_2 + b_{i2}) \\ ke_i &= exp(\beta_3) \end{aligned}$$

where the βs denote fixed-effects parameters and the b_is denote random-effects parameters with an unknown covariance matrix.

For the simulated data, we have considered 10 subjects and the dose 'D' is fixed at 4.5. We have considered the following time stamps for recording the observed concentration of the drug:

t = (0, 0.25, 0.5, 1, 2, 3.5, 5, 7, 10, 12, 15, 18, 21, 24).

(Simulated data can be found in the Appendix as Dataset 1.1)

For the simulated data, time *t* will be considered as the fixed variable and *subject* will be considered as the random variable.

The general model used to build the population and the subject model is:

$$Conc = \beta_0 + \beta_1 Time_i + Subject_j$$

Here, i = 1, 2, 3, …., 14 and j = 1, 2, 3, …., 10

3.2 Incorporation of missing data points

The next step is to create the missing data points in the simulated data.

The clinical trial measures the serum concentration of the drug Theophylline in the body after oral administration to the subject. The observations at the initial and concluding timestamps might be missing due to absence of measurable drug concentration in the subject's body. To maintain the authenticity of the data, we are going to create missing points at both ends of the curve. There will be missing data in the absorption and the elimination phase.

Three cases have been considered so far by removing data points randomly from the data.

1. Two data points missing from the entire dataset.
2. Ten data points missing from the dataset.
3. Fourteen data points missing from the dataset.

Random number of data points missing in the dataset incorporating randomness for number of data points, time and subjects.

3.3 Building statistical models for prediction

Linear mixed models are built to predict the missing data points.

For the linear mixed modelling the data is spliced till the time stamp 5 to obtain spliced curves that are linear to form two subsets. The first subset, Subset 1 represents the absorption phase while the second subset, Subset 2 represents the elimination phase. The third subset is created so that the curve of the subsetted data represents a quadratic curve.

Population models using fixed and random effects and subject-specific models are built for each case given in the previous section.

In the population model with Time and Subject we have the response variable as Concentration, Time as the fixed factor and Subject as the random factor. In the population model with Time only, we have the Concentration as the response variable and Time as the fixed factor. We do not consider the Subject variability in this model. For the subject model, we consider each of the Subjects to be individual data and we fit a model to each individual subject using Time. Here, we have Concentration as the response variable and Time as the fixed variable.

Each dataset with missing data is There are three models for each of the three datasets with missing data. In all there are 27 models for comparison.

Details of the models are given below:

1. Population model with time as the fixed effect and subject as the random effect
2. Population model with time as the fixed effect
3. Subject-wise models with time as the fixed effect

3.4 Computation of Sums of Squares of Imputation and Imputation R^2

To assess the performance of the statistical models for predicting the missing data points, goodness of fit measures are computed.

The Sums of Squares of Imputation for each model and the Imputed R^2 derived from the SSI are used for comparison.

Section 4

Literature Review

4.1. Remedies for missing data in clinical Trials

The missing data in clinical trials can cause various implications during the analysis. As discussed by Guan and Yusoff (2011), it can invalidate hypothesis decisions, lead to unwanted bias in data and significantly underestimate the variability.
The most common strategy of removing the subjects from the analysis might not be an appropriate solution. We might lose out on significant information if these subjects are removed from the study altogether. However, we also maintain the authenticity of the data.
The other common methods include creating a dummy variable for indication of missing data and imputing the missing data points with an arbitrary value.

Some of the traditional approaches practiced widely are discussed below:

4.1.1 Complete Case Analysis

In this method, the subjects with the missing data are excluded from the analysis. This is the most common method in practice and is a default setting in most of the statistical analysis packages. This approach is considered as conservative as the authenticity of the data is maintained.
On the other hand, according to Acock (2005), it might lead to 20%-50% loss of data due to deletion. This can significantly impact the level of significance and the power of the tests. If the sample of the study is not large enough, it can inflate the standard errors and create bias. The complete cases remaining in the study might not be the true representation of the population and might lead to underestimation or overestimation of some effects due to the bias.
If the sample of the study is large enough, the missing points are completely random and power is not a concern then Complete Cases Analysis is a reasonable approach.

4.1.2 Mean Substitution

In this method, the average of the variable is substituted for the missing point in the data.
One of the main concerns of this method is the potential bias. If we have a normal population then substituting the missing point by the mean might be a reasonable strategy. However, if we have a growth curve pattern for our response then replacing an initial point or one of the end points with the mean is not ideal.
Secondly, depending on the amount of the missing values in the data, substitution by the mean can impact the variance significantly Acock (2005). Say, we have 25% of the data that was missing replaced by the average. The deviation of these values from the variable average would be 0, deflating the variance of the variable significantly. Depending on number of missing values for each variable, the attenuation in the variance of the variables will be different. This might result in overestimation or underestimation of effects in analysis resulting in invalid conclusions.

4.1.3 Min-Max Substitution

In this method, the minimum value of the variable is substituted for a missing point on the lower end and the maximum value of the variable is substituted for a missing point on the upper end.
The Min-Max Substitution is suitable for imputation of missing data points either in the initiation or the terminal phase of the trial. This method is also suitable when the response shows a

growth curve and has distinct absorption and elimination phases. Substituting the minimum and maximum values depending on the position is suitable than substituting by the average which would fall in the middle of the curve. This would rectify the attenuation of variance in the variables due to substitution of average to some extent.
However, depending on the number of missing values in the data, this method can still create bias as the minimum and the maximum values used for substitution might not represent the true population.

4.1.4 Missing Indicator Method

In this method, a dummy variable is created to indicate a missing data point for a variable. The dummy variable is coded as '1' if the response variable is missing. The original missing response values are coded as 0 or substituted by an arbitrary value such as the mean. The dummy variable is included in the final statistical model.
The strategy was made popular by Cohen and Cohen (1983) and Cohen, Cohen, West and Aiken (2003). The benefit of this method is maintaining the sample size of the study by not eliminating the incomplete cases as well as maintaining the authenticity of the study to a certain extent.
This method with only one dummy variable will have the same regression estimates as the Complete Cases Analysis, however the indicator variable will indicate the significance of the deviation of the missing points on the average response. Multiple dummy variables for multiple predictors can create multi-collinearity issues in the analysis that can result in bias.
Lastly, the additional dummy variables will utilize some of the degrees of freedom. However, this loss in degrees of freedom might still have less impact than having reduced power by eliminating the missing cases altogether.

In the following section, we are going to use statistical models to predict the missing data points rather than using an arbitrary value. The strategy is start with a basic linear mixed effects model and then proceed to more complex statistical models.

Section 5

Research Methods and Outcome

5.1 Descriptive Statistics

5.1.1 Theophylline Concentration versus Time

The plot (Figure 5.1) shows the relation between Drug Concentration(mg/L) and Time(hrs).

Plot of Theophylline Concentration versus Time by Subject

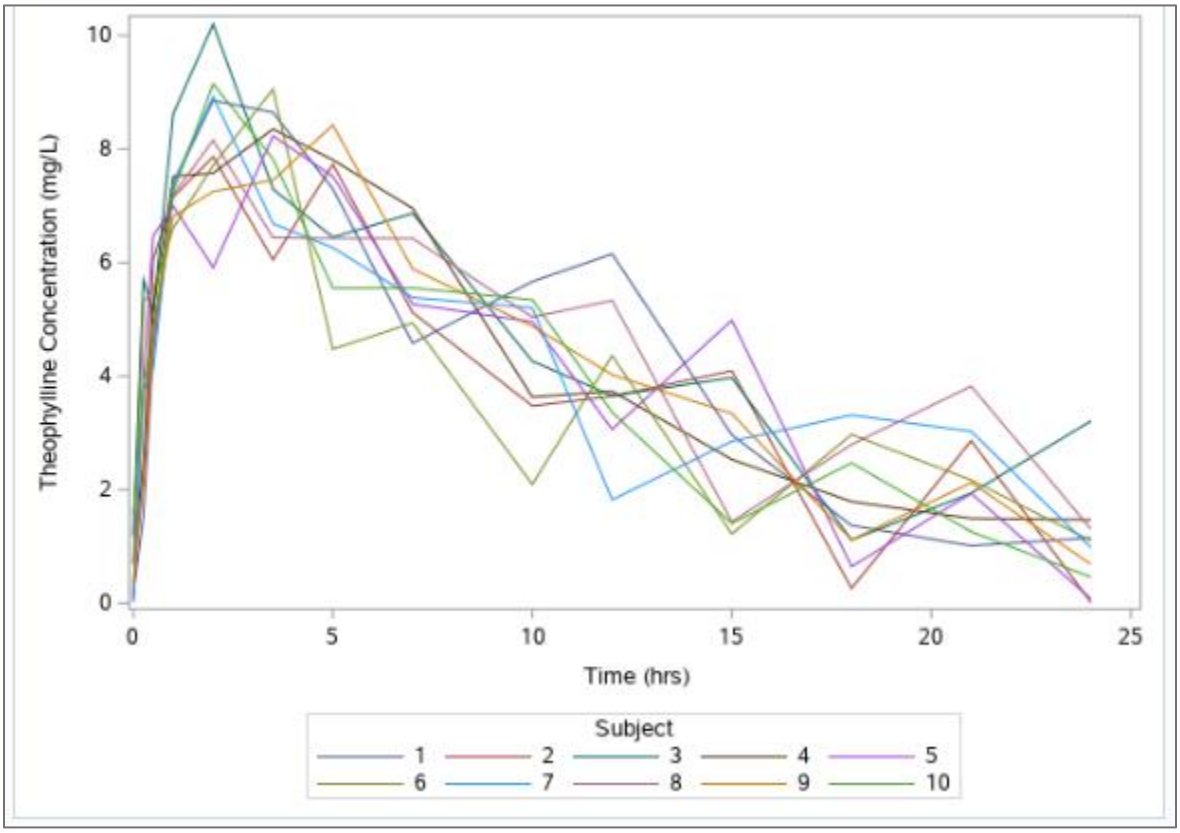

Figure 5.1

Looking at the above plot we can observe that:

1. The concentration-time curves have almost the same shape for all the subjects
2. The rise, peak and decay vary across subjects which might indicate inter-subject variation. In other words, the rate of absorption and elimination for the drug is different for each subject.

5.1.2 Theophylline Concentration versus Time by Subject

Figures (5.2 to 5.6) are individual subject plots for Theophylline Concentration versus Time. The separate curves for subjects are used to check the variability in the Concentration due to Subjects.

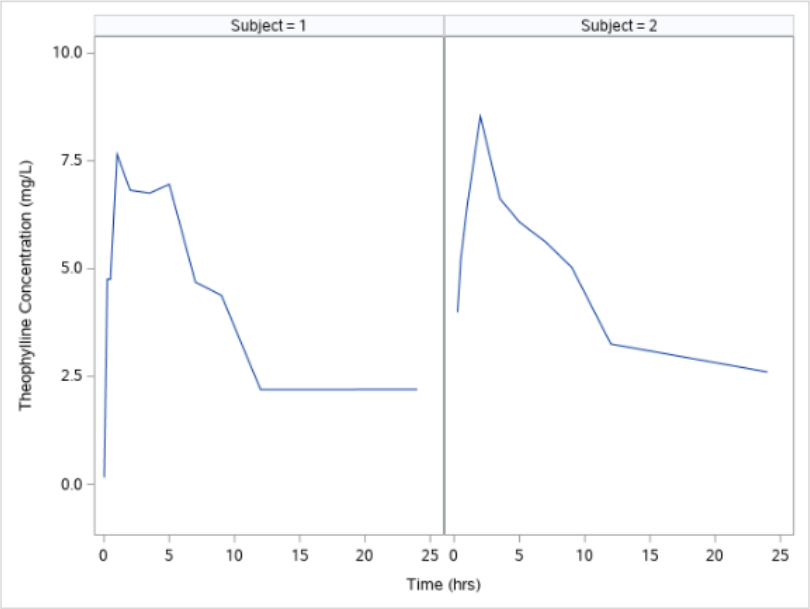

Figure 5.2

Figure 5.2 is the plot of Theophylline Concentration (mg/L) versus Time (hrs) for Subject 1 and Subject 2.

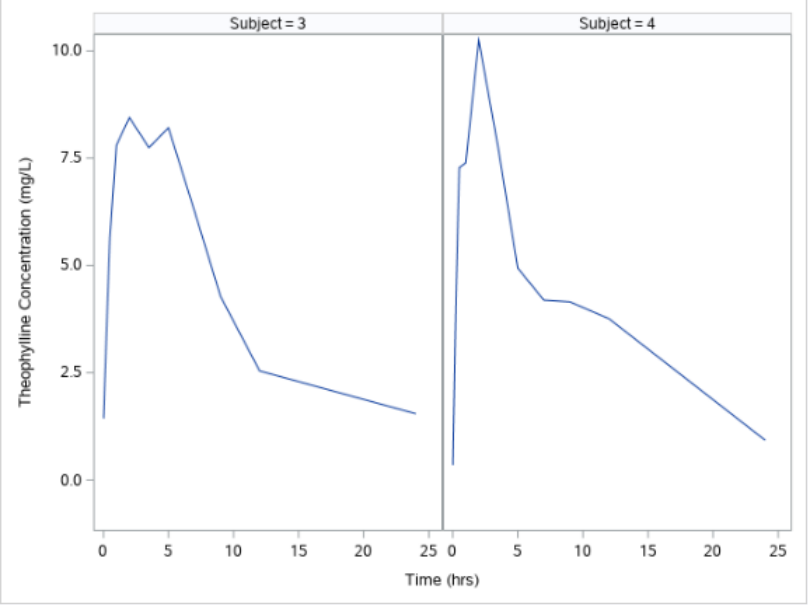

Figure 5.3

Figure 5.3 is the plot of Theophylline Concentration (mg/L) versus Time (hrs) for Subject 3 and Subject 4.

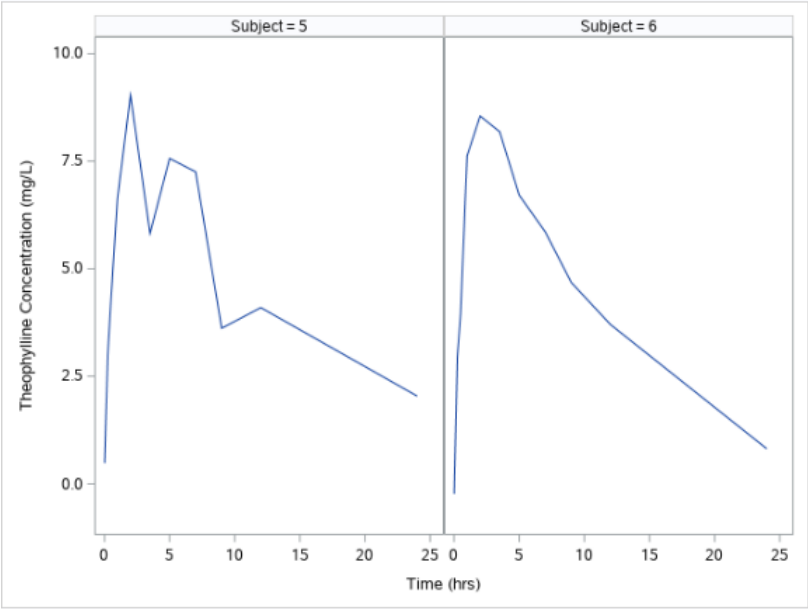

Figure 5.4

Figure 5.4 is the plot of Theophylline Concentration (mg/L) versus Time (hrs) for Subject 5 and Subject 6.

Subject 5 curve shows variation during the elimination phase whereas Subject 6 has a smoother and sharper slope during the elimination phase.

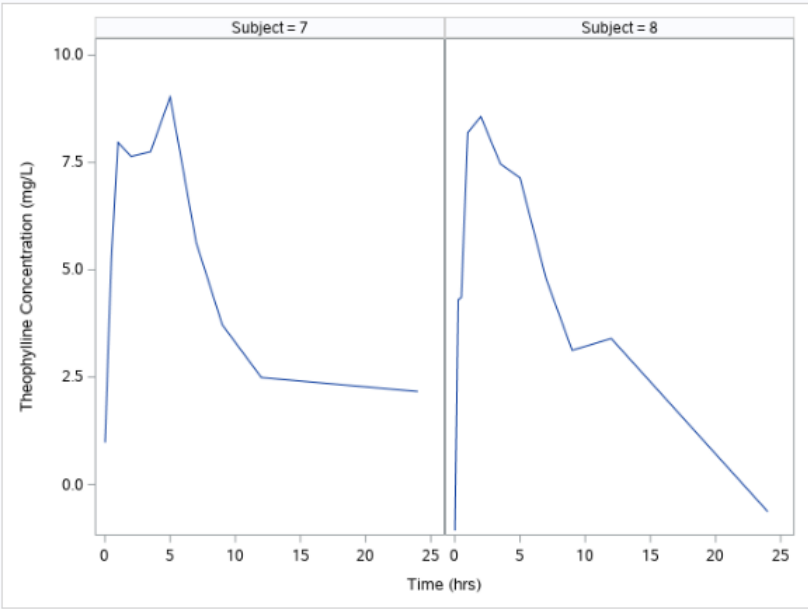

Figure 5.5

Figure 5.5 is the plot of Theophylline Concentration (mg/L) versus Time (hrs) for Subject 7 and Subject 8.

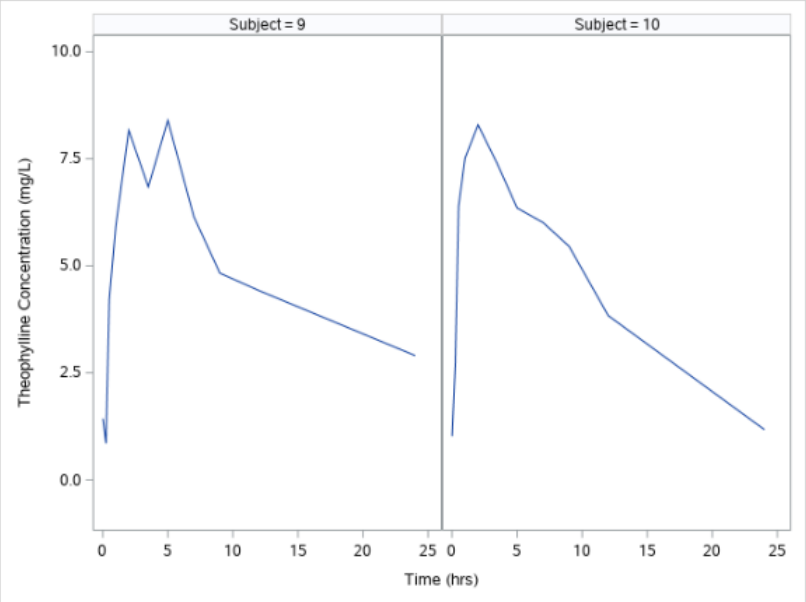

Figure 5.6

Figure 5.6 is the plot of Theophylline Concentration (mg/L) versus Time (hrs) for Subject 9 and Subject 10.

Subject 9 curve shows a bimodal curve and a slower elimination phase. Subject 10 shows a smoother elimination phase.

Looking at the 12 individual plots for each subject, the overall pattern seems to be similar. The concentration of the drug Theophylline in the blood increases during the absorption phase. It decreases again in the elimination phase. The individual plots show that there is a clear variance between the concentration levels which might not be attributed to random error.

5.1.3 Distribution of Theophylline Concentration by Subject

Boxplots of Theophyllne Concentration for each Subject

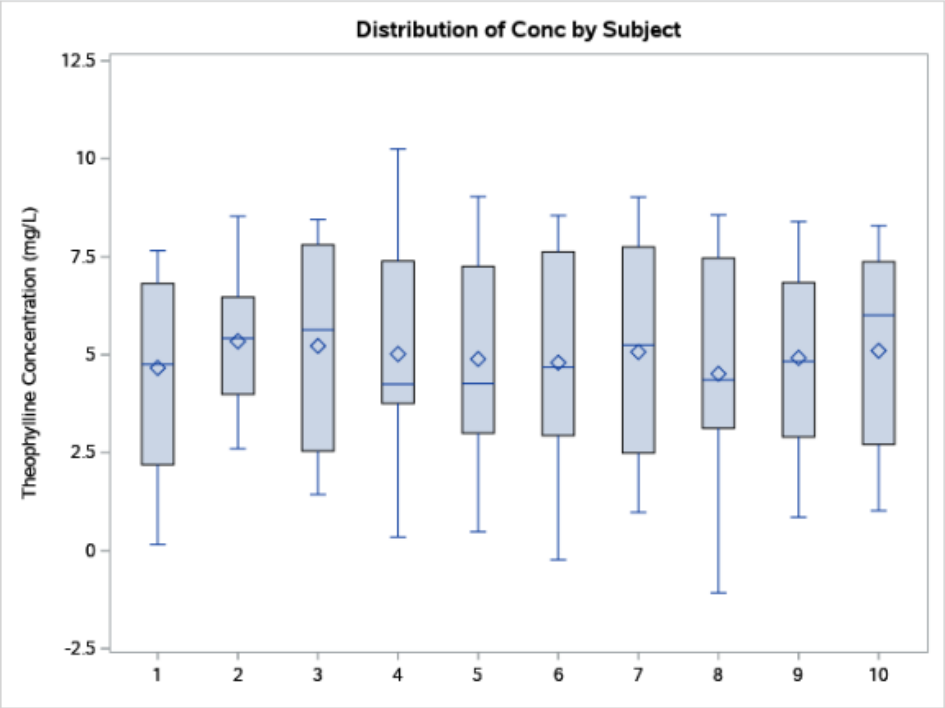

Figure 5.7

Figure 5.7 shows the variability in the Subjects.

The boxplots for each subject are different in size with unequal whisker lengths. This indicates variability in the Theophylline concentration range for each subject. The average Theophylline concentration for the subjects ranges between 4 and 6 mg/L. The lowest and the highest concentrations of the drug for the subjects shows quite a variability. This indicates significant difference in the response variable for each Subject.

The next step is to build statistical models to impute the missing data points in the data. The models will be assessed using various goodness of fit measures.

5.2 Statistical Models, Outcomes and Comparison

There are four cases considered for the imputation and comparison. The data for each of the cases is the same but the number and the positions of the missing data points is selected randomly. The details of the cases are given below:

- Case 1: Two data points missing from the entire dataset. First missing data point is for Subject 2 at time 0; subject is selected at random using R. Second missing data point is for Subject 8 at time 24; subject is selected at random using R.

- Case 2: Missing ten data points are for six subjects. Response for Time 0 is missing for Subjects 4, 7, 9. Response for Time 0.25 is missing for Subjects 4 and 9. Response for Time 21 is missing for Subjects 2 and 5. Response for Time 24 is missing for Subjects 2, 5 and 8; subjects selected at random using random number generator in R.

- Case 3: Missing fourteen data points are for six subjects. Response for Time 0 and 0.25 is missing for Subjects 2, 6, 10. Response for Time 0.5 is missing for Subject 6. Response for Time 18 is missing for Subject 4. Response for Time 21 and 24 is missing for Subjects 4, 8 and 9; subjects selected at random using random number generator in R.

Each case of the simulated data is spliced to form three subsets.
The first two subsets divide the data into two parts to obtain approximate linear curves. Observing the individual curves of all the subjects given in Fig. 5.2 to Fig. 5.6, we can conclude that the absorption phase of most of the subjects ends at 5 hrs and the elimination phase begins and goes on till 24 hrs. We splice the data to obtain approximately linear curves for each of the absorption phase and elimination phase. The first subset encompasses the absorption phase while the second subset encompasses the elimination phase.

The graphs also indicate a possible quadratic curve with a tail on the right side. If we splice the data till the 10 hrs timestamp, we get an approximate quadratic curve. The third subset consists of the data spliced till the 10 hrs to fit a linear quadratic model.

The details of the subsets are given below:

- Part 1: Subset with variable Time less than equal to 5. (Time <= 5)
  Here, i = 1, 2, 3, …., 7
- Part 2: Subset with variable Time more than 5. (Time > 5)
  Here, i = 8, 9, 10, …., 14
- Part 3: Subset with variable Time less than or equal to 10. (Time <= 10)
  Here, i = 1, 2, 3, …., 10

Notations and formulae:

- N: Sample size

- m: Number of missing points in the data
  N* = N – m

- SSI: Sums of Squares of Imputation

  SSI = $\sum_{i=1}^{N^*}(y'_i - \widehat{y'_i})^2$

  Where, y'_i are the non-imputed data points.

- SST: Sums of Squares of Total

  SST = $\sum_{i=1}^{N*}(y'_i - \overline{y'_i})^2$

  Where, y'_i are the non-imputed data points.

- Degrees of Freedom for SSI and SST = N* - 1 = N – m - 1

- Imputed R^2: R^2 calculated using SSI as the numerator and the SST as the denominator

  Imputed R^2 = $1 - \frac{SSI}{SST}$

Assessment of the models for imputation using the goodness of fit measures:

- Lower the Sums of Squares of Imputation (SSI) of the model, better is the model for imputation.
  The SSI indicates the difference between observed value and predicted value of the response variable for that particular model. We want this error to be as small as possible.

- The closer the value of Imputed R^2 is to 1, better is the model for imputation.
  Higher value of the Imputed R^2 indicates that most of the variability in the response is explained by the model used for imputation. Having least amount of variability due to random noise is better.

Note: The values of SSI, SST and Imputed R^2 are going to be same for the Population Models with Time and Subject and Population Models with Time only. The random effect Subject is not going to change the predicted values and the mean estimates.
The random effect affects the variance of the population.
Variance of the Mixed model with the Random effect Subject is $\sigma^2_{error} + \sigma^2_{subject}$ whereas,
Variance for the Model with only the Fixed effect Time is σ^2_{error}

5.2.1. Case 1 – Two data points missing from the entire dataset

First missing data point is for Subject 2 at time 0; subject is selected at random using R.
Second missing data point is for Subject 8 at time 24; subject is selected at random using R

The Case 1 dataset is used to obtain the three subsets – Subset 1 has the data for Time less than or equal to 5, Subset 2 is the data for Time > 5, Subset 3 has data for Time <=10. Each of the Subsets has one missing data point. There are three models built for each Subset.
The models built for Subset 1 are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 2

are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 3 are Quadratic Population Model with Time & Subject, Quadratic Population Model with Time and Quadratic Subject Models with Time.

The models for each Subset need to be compared for their imputation efficiency. The predicted value of the response variable can be compared to the observed value of the response from the dataset. The difference between these two values is used to build a statistical measure for comparison. The Sums of Squares of Imputation is the Sum of the Squares of the Differences between the Predicted and Observed Response for the non-imputed datapoints. This gives an inclination of how good the model is at prediction of the existing data points. The lower the SSI value, better is the model for imputation.

The Imputed R^2 is the given as the ratio of the Residual Sums of Squares and the Total Sums of Squares. It can be obtained by subtracting the ratio of Sums of Squares of Imputation and Total Sums of Squares from 1. This value indicates how much variance in the response is explained by the imputation model. We want this ratio to be as high as possible. The closer the value is to 1, better the model for imputation.

We use the above-mentioned criterion for selection of the best ft for imputation.

The values of the goodness of fit measures for each of the 9 models for Case 1 are given below. The preferred model for each Subset has been highlighted.

The inference drawn from the Table 5.1 is given below the table.

Table 5.1. Comparison of Linear and Quadratic Population and Subject Models for Case 1:

	Model Description	**M**	**SSI**	**SST**	**D.f.**	**Imputed R^2**	
Linear Models for Subset 1 (Time <= 5)	Population Model with Time and Subject	1	321.3514	486.5986	68	0.339597	
	Population Model with Time	1	321.3514	486.5986	68	0.339597	
	Subject Models with Time	1	297.2188	486.5986	5,6	0.389191	
Linear Models for Subset 2 (Time > 5)	Population Model with Time and Subject	1	77.91204	226.0432	68	0.655322	
	Population Model with Time	1	77.91204	226.0432	68	0.655322	
	Subject Models with Time	1	65.74885	226.0432	5,6	0.709131	

Quadratic Models for Subset 3 (Time <= 10)	Population Model with Time and Subject	1	237.7521	419.0333	88	0.432618	
	Population Model with Time	1	237.7521	419.0333	88	0.432618	
	Subject Models with Time	1	384.9207	419.0333	7,8	0.081408	

We can use the Table 5.1 to draw inference about the models used for imputation for Case 1.
For both the linear subsets, Subset 1 and 2, the SSI of the Subject Models is smaller than the SSI of the Population Models with Time & Subject and the Population Models with Time only.

For Subset 1 and 2, the Imputed R^2 of the Subject Models is higher than the Population Models with Time & Subject and Population Models with Time only.
This indicates that the Subject Models are better at imputation of the missing data points in Case 1 than the Population Models.

For Subset 3, the SSI of the Quadratic Population Models is less than the Quadratic Subject Models. The Imputed R^2 of the Quadratic Subject Model at 8 % is significantly less than the Quadratic Population Models at 43%.
This indicates that the Quadratic Population Models are better at imputing the missing points than the Quadratic Subject Model.

The goodness of fit measures of both the Population Models for the Subset 3 are same. The best model is selected using other criterions of model selection. AIC of both the models is same but BIC of the Population Model with Time & Subject is smaller.
Hence, we will select the Quadratic Population Model with Time & Subject for data imputation.

Overall, the Linear Models perform better than the Quadratic Models for imputation of missing points in Case 1. However, the Imputed R^2 is not close to 1. We ideally want a model which has a ratio more than 0.80.

5.2.2. Case 2 – Ten data points missing from the entire dataset

Missing data points are for six subjects. Response for Time 0 is missing for Subjects 4, 7, 9. Response for Time 0.25 is missing for Subjects 4 and 9. Response for Time 21 is missing for Subjects 2 and 5. Response for Time 24 is missing for Subjects 2, 5 and 8; subjects selected at random using random number generator in R.

The Case 2 dataset is used to obtain the three subsets – Subset 1 has the data for Time less than or equal to 5, Subset 2 is the data for Time > 5, Subset 3 has data for Time <=10. Each of the Subsets has five missing data points. There are three models built for each Subset.
The models built for Subset 1 are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 2 are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 3 are Quadratic Population Model with

Time & Subject, Quadratic Population Model with Time and Quadratic Subject Models with Time.

The models for each Subset need to be compared for their imputation efficiency. The predicted value of the response variable can be compared to the observed value of the response from the dataset. The difference between these two values is used to build a statistical measure for comparison. The Sums of Squares of Imputation is the Sum of the Squares of the Differences between the Predicted and Observed Response for the non-imputed datapoints. This gives an inclination of how good the model is at prediction of the existing data points. The lower the SSI value, better is the model for imputation.

The Imputed R^2 is the given as the ratio of the Residual Sums of Squares and the Total Sums of Squares. It can be obtained by subtracting the ratio of Sums of Squares of Imputation and Total Sums of Squares from 1. This value indicates how much variance in the response is explained by the imputation model. We want this ratio to be as high as possible. The closer the value is to 1, better the model for imputation.

We use the above-mentioned criterion for selection of the best ft for imputation.

The values of the goodness of fit measures for each of the 9 models for Case 2 are given below. The preferred model for each Subset has been highlighted.

The inference drawn from the Table 5.2 is given below the table.

Table 5.2. Comparison of Linear and Quadratic Population and Subject Models for Case 2:

	Model Description	**M**	**SSI**	**SST**	**D.f.**	**Imputed R^2**	
Linear Models for Subset 1 (Time <= 5)	Population Model with Time and Subject	5	282.6106	404.9194	64	0.302057	
	Population Model with Time	5	282.6106	404.9194	64	0.302057	
	Subject Models with Time	5	253.0196	404.9194	4,5,6	0.375136	
Linear Models for Subset 2 (Time > 5)	Population Model with Time and Subject	5	75.00623	204.0997	64	0.632502	
	Population Model with Time	5	75.00623	204.0997	64	0.632502	
	Subject Models with Time	5	61.6037	204.0997	4,5,6	0.698169	
Quadratic Models for Subset 3 (Time <= 10)	Population Model with Time and Subject	5	223.9118	382.9611	84	0.415315	
	Population Model with Time	5	223.9118	382.9611	84	0.415315	
	Subject Models with Time	5	340.2177	382.9611	6,7,8	0.111613	

We can use the Table 5.1 to draw inference about the models used for imputation for Case 1. Case 2 has ten missing data points while Case 1 has two missing data points. Increasing the number of missing points in the dataset has not affected the models' imputation performance significantly.

For both the linear subsets, Subset 1 and 2, the SSI of the Subject Models is smaller than the SSI of the Population Models with Time & Subject and the Population Models with Time only.
For Subset 1 and 2, the Imputed R^2 of the Subject Models is higher than the Population Models with Time & Subject and Population Models with Time only.
This indicates that the Subject Models are better at imputation of the missing data points in Case 1 than the Population Models.

For Subset 3, the SSI of the Quadratic Population Models is less than the Quadratic Subject Models. The Imputed R^2 of the Quadratic Subject Model at 11 % is significantly less than the Quadratic Population Models at 42%.
This indicates that the Quadratic Population Models are better at imputing the missing points than the Quadratic Subject Model.
The goodness of fit measures of both the Population Models for the Subset 3 are same. The best model is selected using other criterions of model selection. AIC of both the models is same but BIC of the Quadratic Population Model with Time & Subject is smaller.
Hence, we will select the Quadratic Population Model with Time & Subject for data imputation.

Overall, the Linear Models perform better than the Quadratic Models for imputation of missing points in Case 2. However, the Imputed R^2 is not close to 1. We ideally want a model which has a ratio more than 0.80.

5.2.3. Case 3 – Fourteen data points missing from the entire dataset

Missing data points are for six subjects. Response for Time 0 and 0.25 is missing for Subjects 2, 6, 10. Response for Time 0.5 is missing for Subject 6. Response for Time 18 is missing for Subject 4. Response for Time 21 and 24 is missing for Subjects 4, 8 and 9; subjects selected at random using random number generator in R.

The Case 3 dataset is used to obtain the three subsets – Subset 1 has the data for Time less than or equal to 5, Subset 2 is the data for Time > 5, Subset 3 has data for Time <=10. Each of the Subsets has seven missing data points. There are three models built for each Subset.
The models built for Subset 1 are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 2 are Linear Population Model with Time & Subject, Linear Population Model with Time and Linear Subject Models with Time. The models built for Subset 3 are Quadratic Population Model with Time & Subject, Quadratic Population Model with Time and Quadratic Subject Models with Time.
The models for each Subset need to be compared for their imputation efficiency. The predicted value of the response variable can be compared to the observed value of the response from the dataset. The difference between these two values is used to build a statistical measure for comparison. The Sums of Squares of Imputation is the Sum of the Squares of the Differences

between the Predicted and Observed Response for the non-imputed datapoints. This gives an inclination of how good the model is at prediction of the existing data points. The lower the SSI value, better is the model for imputation.
The Imputed R^2 is the given as the ratio of the Residual Sums of Squares and the Total Sums of Squares. It can be obtained by subtracting the ratio of Sums of Squares of Imputation and Total Sums of Squares from 1. This value indicates how much variance in the response is explained by the imputation model. We want this ratio to be as high as possible. The closer the value is to 1, better the model for imputation.
We use the above-mentioned criterion for selection of the best ft for imputation.
The values of the goodness of fit measures for each of the 9 models for Case 3 are given below. The preferred model for each Subset has been highlighted.

The inference drawn from the Table 5.2 is given below the table.

Table 5.3. Comparison of Linear and Quadratic Population and Subject Models for Case 3:

	Model Description	**M**	**SSI**	**SST**	**D.f.**	**Imputed R^2**	
Linear Models for Subset 1 (Time <= 5)	Population Model with Time and Subject	7	297.4267	429.6015	62	0.307668	
	Population Model with Time	7	297.4267	429.6015	62	0.307668	
	Subject Models with Time	7	240.4443	429.6015	3,4,6	0.440309	
Linear Models for Subset 2 (Time > 5)	Population Model with Time and Subject	7	71.76221	209.9604	62	0.658211	
	Population Model with Time	7	71.76221	209.9604	62	0.658211	
	Subject Models with Time	7	56.97727	209.9604	3,4,6	0.728628	
Quadratic Models for Subset 3 (Time <= 10)	Population Model with Time and Subject	7	199.8548	323.2613	82	0.381755	
	Population Model with Time	7	199.8548	323.2613	82	0.381755	
	Subject Models with Time	7	280.5181	323.2613	5,6,8	0.132225	

We can use the Table 5.1 to draw inference about the models used for imputation for Case 3.

Case 3 has fourteen missing data points while Case 2 has ten missing data points. Increasing the number of missing points in the dataset has not affected the models' imputation performance significantly.
For both the linear subsets, Subset 1 and 2, the SSI of the Subject Models is smaller than the SSI of the Population Models with Time & Subject and the Population Models with Time only.
For Subset 1 and 2, the Imputed R^2 of the Subject Models is higher than the Population Models with Time & Subject and Population Models with Time only.
This indicates that the Subject Models are better at imputation of the missing data points in Case 3 than the Population Models.

For Subset 3, the SSI of the Population Models is less than the Subject Models.
The Imputed R^2 of the Subject Model at 13 % is significantly less than the Population Models at 38%.
This indicates that the Quadratic Population Models are better at imputing the missing points than the Quadratic Subject Models.
The goodness of fit measures of both the Population Models for the Subset 3 are same. The best model is selected using other criterions of model selection. AIC of both the models is same but BIC of the Population Model with Time & Subject is smaller.
Hence, we will select the Quadratic Population Model with Time & Subject for data imputation.

Overall, the Linear Models perform better than the Quadratic Models for imputation of missing points in Case 2. However, the Imputed R^2 is not close to 1. We ideally want a model which has a ratio more than 0.80.

5.3 Discussion

From the Theophylline Concentration versus Time plot given in Figure 5.1, it can be concluded that the concentration of the drug in the body does not have a linear trend. A polynomial term in the linear model will fit the data better than a linear model without any polynomial terms. It is also easier and more straightforward than fitting a complicated non-linear model to the data.

The linear Subject models for Subsets 1 and 2, that is the absorption phase and the elimination phase in all the three cases perform better than the Population models with Subject and Population models with Time only. The distribution of the Theophylline concentration in the body for the subjects varies significantly as seen in the boxplot Figure 5.7. This indicates that each subject has a different rate of absorption and elimination of the drug from the body. This suggests that an individual model for each Subject might be a better fit than a population model including all the Subjects. This is supported by the results from all three cases where the SSI of the subject models is lower than the population models.

The Subset 3 for all three cases indicate a curve. More the data points in the dataset, better will the fit of the curve. Getting a good fit of a subject model for each subject becomes difficult if the subject has higher number of missing points. In such a case, it is better to fit a population model. This is supported by the analysis as well. The population models for all the three cases perform better than the subject

models. The SSI of the quadratic population models of Subset 3 of the three cases is less the quadratic subject models.

The subject models will perform better when there are substantial data points to fit a good model for each subject. In cases where there are not enough data points for each subject, population models will give a better estimate for imputation.

The imputation process's primary objective is to enable the use of most of the available data by imputing the missing values in the data while maintaining the authenticity. It is crucial that we include only the essential covariates in the model used for imputation. This can be decided based on the nature of the data and the number of covariates in the data.

A subset of core covariates can be considered from all of the available covariates for the model for imputation. The significant covariates from this subset can be obtained by model selection techniques such as forward, backward selection etc. The level of significance for decision of inclusion or exclusion of the covariate in the model of imputation does not need to be as stringent as that while implementing model selection for data analysis.

The additional covariates from the data which were not included in the above-mentioned subset, can be included in the model for analysis using the usual model selection techniques. Usually the covariates included in the model for imputation will be included in the model for analysis along with other significant covariates.

We do not want the model for imputation to be the same as the model used for the analysis. Having the same model for imputing the missing values and analysis will lead to overfitting of the data. Analyzing the data which has missing values replaced by the same model will fit better to the data leading to incorrect results. It is crucial to choose an appropriate model for imputation to avoid these issues.

The recommendations and scope of this study for practical applications are discussed in detail in the next chapter along with the shortcomings.

Section 6

Conclusion and Scope

6.1. Conclusion

First, the performance of the statistical models at imputing the missing data points starts decreasing as the number of missing points in the data increase.
Secondly, the Sums of Squares of Imputation behave as expected. A better model of imputation will have smaller Sums of Squares of Imputation. The Imputed R^2 goodness of fit measure is a simple tool to assess how good the statistical model of imputation is for predicting the missing data. It gives a measure of how much of the variability in the response is explained by the model of imputation.
Lastly, the linear models are better than the quadratic models at imputation for the simulated data. Among the linear models, the population models with the fixed effect Time and random effect Subject included is a better than the population model without the random effect Subject.

6.2. Shortcomings

We have a simple dataset with one fixed effect and one random effect. Clinical data, in real life, is always going to be more complex with multiple covariates and a large sample size.
The curve of Concentration with Time indicates some type of a growth curve. A non-linear model with mixed effects might be a better fit for the data. The linear and quadratic models used for imputation for the simulated data are not the best models for the given data. There might be better complex models which might predict the missing points better.

6.3. Recommendations

Few recommendations to improve the study and to make it more applicable to generic cases are present.
Firstly, incorporating multiple covariates such as Age, Gender, Weight, Metabolism Rate, Blood Routine observations, etc. in the dataset to assess the performance of the models with multiple dimensions. Increasing the sample size of the study to reduce random error.
Secondly, assessing the performance of the goodness of fit measures, SSI and Imputed R^2 when they have deal with collinearity of multiple covariates.
Thirdly, fitting statistical models with higher complexity to fit curve of the data. Non-linear mixed effects model such as One-compartment Pharmacokinetic model might be a better fit for the data.
Lastly, increasing the sample size of the study as real-life trial would have a much larger sample size.

6.4. Scope

Incorporating multiple covariates in the data to assess the performance of the goodness of fit criterion for imputation for complex models. The Concentration Time curve indicates a growth curve patter to the data. Fitting non-linear growth curves to the data will be an ideal approach.

Applying this method to a real-life data instead of a simulated data is future goal.
The goodness of fit measure Imputed R^2 might have to be modified when there are multiple covariates in the data. The measure might have to be penalized to incorporate the multiple variables in the model.
The Sums of Squares of Imputation theoretically, should follow Chi-square distribution. A test statistic based on the Sums of Squares of Imputation and the degrees of freedom can be built. This test statistic theoretically should follow F-distribution. This would help in deciding whether the model is useful for imputation. Validating this hypothesis is one of the future objectives.

References

Acock A. C. (2005). Working with missing values. Journal of Marriage and the Family, 67(4), 1012–1028. 10.1111/j.1741-3737.2005.00191.x

Berg, P., McConnell, E.W., Hicks, L.M. et al. Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. BMC Bioinformatics 20, 102 (2019) doi:10.1186/s12859-019-2619-6

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Guan NC, Yusoff MS. Missing values in data analysis: Ignore or impute? Educ Med J. 2011;3:e6–11

Hollestein L., Carpenter J. (2017) Missing data in clinical research: an integrated approach. Br J Dermatol. 2017; 177: 1463-1465

Ibrahim, J. G., Chu, H., & Chen, M. H. (2012). Missing data in clinical studies: issues and methods. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 30(26), 3297–3303. doi:10.1200/JCO.2011.38.7589

Pinheiro, J. C. and Bates, D. M. (1995), “Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model,” Journal of Computational and Graphical Statistics, 4, 12–35.

Richardson-Harman N, Parody R, Anton P, McGowan I, Doncel G, Thurman AR, et al. Analytical Advances in the *Ex Vivo* Challenge Efficacy Assay. AIDS Res Hum Retroviruses. 2016:Epub ahead of print Dec. 16.

Ross, Zoe Michelle, "Statistical linear mixed models for evaluation of training program in hand surgery chief residents" (2017). Honors Theses.

Appendix

Dataset 1.1

Obs. No.	Subject	Time	Dose	Conc
1	1	0	4.5	0.166926
2	1	0.25	4.5	1.543568
3	1	0.5	4.5	4.517274
4	1	1	4.5	7.381421
5	1	2	4.5	8.859213
6	1	3.5	4.5	8.645822
7	1	5	4.5	7.300276
8	1	7	4.5	4.579653
9	1	10	4.5	5.666462
10	1	12	4.5	6.152977
11	1	15	4.5	2.963063
12	1	18	4.5	1.368785
13	1	21	4.5	1.010404
14	1	24	4.5	1.150451
15	2	0	4.5	0.675363
16	2	0.25	4.5	2.559472
17	2	0.5	4.5	6.061638
18	2	1	4.5	7.158622
19	2	2	4.5	7.865775
20	2	3.5	4.5	6.048914
21	2	5	4.5	7.731559
22	2	7	4.5	5.120209
23	2	10	4.5	3.474936
24	2	12	4.5	3.645725
25	2	15	4.5	4.093258
26	2	18	4.5	0.26015
27	2	21	4.5	2.862773
28	2	24	4.5	0.012882
29	3	0	4.5	1.227623
30	3	0.25	4.5	5.699911
31	3	0.5	4.5	5.141344
32	3	1	4.5	8.626037
33	3	2	4.5	10.20318
34	3	3.5	4.5	7.2956
35	3	5	4.5	6.446552
36	3	7	4.5	6.86881
37	3	10	4.5	4.263719

38	3	12	4.5	3.654819
39	3	15	4.5	3.971287
40	3	18	4.5	1.11396
41	3	21	4.5	1.937399
42	3	24	4.5	3.208905
43	4	0	4.5	1.168756
44	4	0.25	4.5	2.283667
45	4	0.5	4.5	5.206775
46	4	1	4.5	7.522831
47	4	2	4.5	7.57764
48	4	3.5	4.5	8.353213
49	4	5	4.5	7.809229
50	4	7	4.5	6.955636
51	4	10	4.5	3.631474
52	4	12	4.5	3.732407
53	4	15	4.5	2.531836
54	4	18	4.5	1.789414
55	4	21	4.5	1.492662
56	4	24	4.5	1.468216
57	5	0	4.5	0.015852
58	5	0.25	4.5	4.088644
59	5	0.5	4.5	6.478556
60	5	1	4.5	6.995201
61	5	2	4.5	5.914913
62	5	3.5	4.5	8.23369
63	5	5	4.5	7.519846
64	5	7	4.5	5.268065
65	5	10	4.5	4.959453
66	5	12	4.5	3.063056
67	5	15	4.5	4.982898
68	5	18	4.5	0.644385
69	5	21	4.5	1.934162
70	5	24	4.5	0.076504
71	6	0	4.5	0.720138
72	6	0.25	4.5	5.309323
73	6	0.5	4.5	5.523851
74	6	1	4.5	6.614434
75	6	2	4.5	7.715629
76	6	3.5	4.5	9.055209
77	6	5	4.5	4.475876
78	6	7	4.5	4.937212

79	6	10	4.5	2.09295
80	6	12	4.5	4.362354
81	6	15	4.5	1.22023
82	6	18	4.5	2.972093
83	6	21	4.5	2.168738
84	6	24	4.5	1.098148
85	7	0	4.5	0.06548
86	7	0.25	4.5	3.752712
87	7	0.5	4.5	4.135596
88	7	1	4.5	7.415654
89	7	2	4.5	8.923245
90	7	3.5	4.5	6.687456
91	7	5	4.5	6.264115
92	7	7	4.5	5.385699
93	7	10	4.5	5.20188
94	7	12	4.5	1.821922
95	7	15	4.5	2.85194
96	7	18	4.5	3.313781
97	7	21	4.5	3.026171
98	7	24	4.5	0.978879
99	8	0	4.5	0.391919
100	8	0.25	4.5	2.160716
101	8	0.5	4.5	4.527807
102	8	1	4.5	7.198262
103	8	2	4.5	8.157101
104	8	3.5	4.5	6.439462
105	8	5	4.5	6.425069
106	8	7	4.5	6.426851
107	8	10	4.5	5.038696
108	8	12	4.5	5.333265
109	8	15	4.5	1.434062
110	8	18	4.5	2.794701
111	8	21	4.5	3.8224
112	8	24	4.5	1.304759
113	9	0	4.5	0.316705
114	9	0.25	4.5	2.292128
115	9	0.5	4.5	4.872779
116	9	1	4.5	6.819063
117	9	2	4.5	7.256018
118	9	3.5	4.5	7.454822
119	9	5	4.5	8.424974

120	9	7	4.5	5.887813
121	9	10	4.5	4.880832
122	9	12	4.5	4.024156
123	9	15	4.5	3.340954
124	9	18	4.5	1.113572
125	9	21	4.5	2.122453
126	9	24	4.5	0.686172
127	10	0	4.5	1.556844
128	10	0.25	4.5	3.826569
129	10	0.5	4.5	4.834146
130	10	1	4.5	7.261497
131	10	2	4.5	9.149742
132	10	3.5	4.5	7.828541
133	10	5	4.5	5.554357
134	10	7	4.5	5.554271
135	10	10	4.5	5.346518
136	10	12	4.5	3.362648
137	10	15	4.5	1.403637
138	10	18	4.5	2.469548
139	10	21	4.5	1.257984
140	10	24	4.5	0.460307